DOCUMENT RESUME

ED 408 525 CG 027 770

AUTHOR Forsyth, G. Alfred; And Others
TITLE Developing and Assessing Students' Abilities To Interpret Research.
PUB DATE Aug 95
NOTE 29p.; Paper presented at the Annual Meeting of the American Psychological Association (103rd, New York, NY, August, 1995). For related document, see CG 027 769.
PUB TYPE Information Analyses (070) -- Speeches/Meeting Papers (150) -- Tests/Questionnaires (160)
EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Data Interpretation; Evaluation Methods; Evaluation Research; *Evaluative Thinking; Higher Education; *Interpretive Skills; Professional Education; Research Skills; Statistical Analysis; *Student Evaluation; Thinking Skills
IDENTIFIERS *Vignettes

ABSTRACT

A recent conference on statistics education recommended that more emphasis be placed on the interpretation of research (IOR). Ways for developing and assessing IOR and providing a systematic framework for creating and selecting instructional materials for the independent assessment of specific IOR concepts are the focus of this paper. The recommended assessment procedure to evaluate IOR abilities consists both of vignettes (research-report summaries) and questions designed to assess the students' interpretations of those vignettes. Vignettes could be selected to systematically vary on different features, such as random sampling as opposed to using an available group of subjects. A systematic framework is introduced here, consisting of 4 features, that can be used to write sets of vignettes for each cell of a 16-cell taxonomy. The features are: (1) random assignment versus classificatory independent variable; (2) a dependent variable that is either life-experience meaningful or not; (3) results that are counter to popular beliefs or for which there is no clear expectation of outcome; and (4) the independent variable having levels that are quantitatively different versus the independent variable not having an underlying continuum. A booklet containing a sample vignette for each cell of the taxonomy is appended. (RJM)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 408 525

DEVELOPING AND ASSESSING STUDENTS' ABILITIES TO INTERPRET RESEARCH

G. Alfred Forsyth
Millersville University
Millersville, PA 17551

Peter H. Bohling
Bloomsburg University
Bloomsburg, PA 17815

T. William Altermatt
University of Illinois
Champaign, Illinois

Presented at the Annual Meetings of the American Psychological Association, August 1995, New York, N.Y.

BEST COPY AVAILABLE

PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

A. Forsyth

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☐ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

CG027770

DEVELOPING AND ASSESSING STUDENTS' ABILITIES TO INTERPRET RESEARCH

G. Alfred Forsyth
(Millersville University)

Peter H. Bohling
(Bloomsburg University)

T. William Altermatt
(University of Illinois)

One recommendation resulting from a recent conference on statistics education was that statistics courses should place a heavier emphasis on the interpretation of research (Hogg, 1991). The National Science Foundation Project 2061 (1992) indicated that the development of an ability to apply statistical knowledge has not kept pace with either rote memory or calculation knowledge of statistics. The importance of developing interpretation-of-research abilities was also recognized in the AAAS Benchmarks for Science Literacy (1993).

The extensive attention to factual knowledge and computational procedures at the expense of developing interpretation skills in statistics courses may account for the inability of students to interpret research correctly. Many statistics courses do not require students to interpret reports of research. Students learn to develop those skills and abilities that they know will be assessed. If teachers of statistics and research methods courses have the development of interpretation skills as a course goal, they must provide students with interpretation-of-research exercises and must assess those skills.

Several factors contribute to the lack of attention to developing and assessing interpretation-of-research abilities. First, reading a complete research report is time intensive. A second impediment is the challenge for teachers to find research articles that differ systematically in reported features such as random sampling vs. available groups, random vs. classificatory assignment of subjects, number of subjects per group, p-value, and levels of strength-of-relationship indices. Finally, media reports of research usually do not provide sufficient information for students to draw appropriate conclusions.

One purpose of this presentation is to provide a procedure for developing and assessing interpretation-of-research abilities. A second purpose is to provide a systematic framework for developing and selecting instructional materials for the independent assessment of specific interpretation-of-research concepts.

The recommended assessment procedure to evaluate interpretation-of-research abilities consists of research report vignettes along with questions designed to assess the students' interpretations of those vignettes. The following questions would be answered by the students as they interpret each vignette:

1. What are the independent and dependent variables?
2. Was any systematic relationship found between the independent and dependent variables?

3. How confident would you be in drawing cause-and-effect conclusions?
4. To what extent can the results of the study be generalized to individuals other than those in the study?
5. How strong is the relationship between the independent and dependent variables?
6. How important do you consider the relationship between variables to be?
7. What additional information should have been provided to permit a clearer interpretation of the research?

The assessment of a student's interpretation skills requires an examination of responses across vignettes. For example, suppose that a student responded to questions for eight vignettes in which four involved random assignment of subjects to groups and four in which the independent variable was classificatory. If a student understands that random assignment leads to more confidence in drawing cause-and-effect conclusions, the average cause-and-effect confidence ratings for the four random-assignment vignettes should be higher than the average cause-and-effect confidence ratings for the four classificatory-study vignettes. The eight vignettes would vary in other features unrelated to drawing cause-and-effect conclusions.

A student's understanding of the relationship between random-sampling and generalizability could be examined with these

same eight vignettes if four involved random sampling and four used available (convenient) groups. The average confidence in generalizing should be higher for the four random-sampling vignettes than for the four available-group vignettes. The specific information in the vignettes and the related interpretation questions would be varied based on what interpretation skills a teacher wishes to assess.

This assessment procedure permits a teacher to grade a student on each of several aspects of interpretation. This makes the assessment process helpful for diagnostic as well as for grading purposes. The assessment procedure also allows a teacher to determine the degree to which a class understands specific aspects of interpreting reports of research. Faculty interested in value-added, outcome-based assessment might present vignettes both at the beginning and end of a course or portion of a course.

Depending on the specific aspects of research interpretation to be studied, vignettes could be selected to systematically vary on one or more of the following features:

1) the independent variable having levels that are quantitatively different vs. the independent variable not having an underlying continuum;
2) random sampling vs. using an available group of subjects;
3) random assignment vs. classificatory grouping of subjects;

4) number of subjects in the study;

5) the dependent variable having life experience meaningfulness (e.g., grade in a course) or not (e.g., score on an unpublished emotional empathy scale);

6) p values varying from .05 to .0001;

7) eta-squared or r-squared as small, medium, or large;

8) confidence intervals for differences between means with lower limits varying in distance from zero;

9) results that are congruent with popular beliefs vs. results that are counter to popular beliefs or for which there are not clear expectations about a relationship.

This last feature was included because of research that indicates that judgments about independent-dependent variable relationships are based on students' initial beliefs about the relationship rather than on the research methods used in the study (e.g., Forsyth, Bohling and May, 1991).

The systematic framework that we have used is to write sets of vignettes for each cell of a 16-cell taxonomy created by crossing four of the above nine features. Specifically, these four features are: 1) random assignment vs. classificatory independent variable, 2) a dependent variable that is life-experience meaningful or not, 3) results that are congruent with popular beliefs vs. results that are counter to popular beliefs or for which there is no clear expectation of outcome, and

4) the independent variable having levels that are quantitatively different vs. the independent variable not having an underlying continuum.

These four features were chosen for the taxonomy because they cannot be varied easily in modifying a vignette. Each vignette within each of the 16-cells comprising the taxonomy can be varied in number of subjects, p-value, random sample vs. available-group, eta-squared and/or r-squared magnitude, and strength-of-effect reflected by the confidence interval. Which of these nine features are varied or held constant depends on the specific interpretation skills the faculty member or researcher wishes to assess. For example, if interested in assessing students' judgments of generalizability and cause-and-effect, an instructor would use a set of vignettes that cross the random-sampling vs. available group feature with the random-assignment vs. classificatory feature.

A booklet containing a sample vignette for each cell of the taxonomy is appended. These vignettes are intended to be used as guides in the development of additional vignettes for each cell. A set of nine questions for each vignette is presented in the booklet. Survey booklets being used in research that assesses specific interpretation-of-research abilities may be obtained from the authors.

References

American Association for the Advancement of Science (1993). Benchmarks for science literacy, Oxford University Press.

Forsyth, G. A., Bohling, P. H., and May, R. B. (1991). Factor affecting the misinterpretation of research. Paper presented at the 99th Annual Meeting of the American Psychological Association, San Francisco, California.

Hogg, R. V. (1991). Statistical education: Improvements are badly needed. The American Statistician, 45, 342-343.

National Science Foundation (1992). Indicators of science and mathematics education. Washington, D.C.

Forsyth, Bohling, Altermatt

SAMPLE VIGNETTE BOOKLET

The number for each of the sample vignettes corresponds to the number in the vignette taxonomy on the following page. The 16 cells are the result of crossing (1) Presence or absence of random assignment, (2) results that confirm or do not confirm popular beliefs, (3) dependent variables for which subjects have a concrete referent or not, and (4) an independent variable that is quantitative or qualitative. The features that can be changed or omitted within each vignette are: (1) number of subjects, (2) p-value, (3) random sample vs available group, (4) eta-squared or r-squared magnitude, and (5) strength of effect reflected by the confidence interval.

After reading each vignette, students answer the nine questions, presented on the page following the vignette taxonomy. Questions 1, 2 and 9 are open-ended. The scale for questions 3, 4, 5, and 6 is:

1	2	3	4	5
not at all confident	somewhat unconfident	neutral	somewhat confident	very confident

---

The scale for question 7 is:

1	2	3	4	5
very weak	weak	neutral	strong	very strong

The scale for question 8 is:

1	2	3	4	5
very unimportant	unimportant	neutral	important	very important

VIGNETTE TAXONOMY

CONFIRMS POPULAR BELIEF

Independent Variable

NO CONCRETE REFERENT	Qualitative	Quantitative
Random Assignment	1	2
Classificatory	3	4

NO POPULAR BELIEF CONFIRMATION

Independent Variable

NO CONCRETE REFERENT	Qualitative	Quantitative
Random Assignment	5	6
Classificatory	7	8

CONFIRMS POPULAR BELIEF

Independent Variable

CONCRETE REFERENT	Qualitative	Quantitative
Random Assignment	9	10
Classificatory	11	12

NO POPULAR BELIEF CONFIRMATION

Independent Variable

CONCRETE REFERENT	Qualitative	Quantitative
Random Assignment	13	14
Classificatory	15	16

Research Methods Features Varied Within Vignettes

1. Random Sample vs. Available Group
2. Eta-Squared or r-Squared Magnitude
3. Confidence Interval Strength of Effect
4. p-Value
5. Number of Subjects

INTERPRETATION-OF-RESEARCH QUESTIONS

1. What is the dependent variable or behavior of interest?

2. What is the independent variable?

3. Based on this research study description, how confident are you that the researcher found a systematic relationship between the independent and dependent variables?

4. Based on the research methods described, how confident would you be in generalizing the results of the study to individuals other than those in the study?

5. Based on the research methods described, how confident would you be in concluding that differences in the independent variable caused differences in the dependent variable?

6. If the proportion of variability in the dependent variable that is accountable by knowing the independent variable was reported with r-squared, how confident would you be in concluding that differences in the independent variable caused differences in the dependent variable?

7. How strong do you consider the relationship to be between the independent and dependent variable?

8. How important do you consider the relationship to be between the independent and dependent variable?

9. What additional information should have been provided to permit a clearer interpretation of the research?

VIGNETTE 1

A human relations psychologist, working for a nation-wide family restaurant, studied the relationship between waiter/waitress training and customer satisfaction. In one study, she decided to use only (female) waitresses to examine if there were differences between three training methods. The psychologist randomly sampled 60 newly-hired waitresses from all the new waitresses at the company's 5,000 restaurants. She randomly assigned 20 of these to the efficiency group. Their training focused on ways to speed the process of taking orders and getting the food to the table. She randomly assigned another 20 new waitresses to the friendly group. Their training emphasized friendliness to customers and ways to make the dining mood a positive one (e.g., cheery comments about how cute the children are or what a nice outfit an adult customer is wearing). The third group of 20 new waitresses are trained to know the menu, how to write down and deliver orders to the kitchen, and how to know when the order is ready.

Twenty-five customers, served by each of these 60 waitresses, are asked to give a dining-experience satisfaction rating. Customers rated their satisfaction completing a questionnaire that asked them to indicate their satisfaction with each of several aspects of their dining experience. The average of these 25 ratings is the score for each waitress.

Analyses of the data by the human-relations psychologist indicated that the friendly waitresses received a higher average satisfaction rating than the efficient waitresses. The efficient waitresses received a higher average satisfaction rating than the basics group. The differences among these three means were larger than expected by chance at the .01 level of significance ($p < .01$). Twenty-one percent of the variability in customer satisfaction ratings was accountable by knowing which type of training the waitresses received (eta-squared = .21). A 99% confidence interval indicated that customer satisfaction would be somewhere between 0.5 to 3.5 higher if the population of waitresses were given friendly rather than basics training.

VIGNETTE 2

The Director of Resident Life at a state university studied the relationship between roommate similarity in attitudes and values and the level of roommate likability. She included in her study the 120 females assigned to a residence hall that is filled exclusively by female freshmen. At freshman orientation in the summer, the director administered an attitudes and values scale to these 120 freshmen women. Based on their answers, the director randomly assigns 40 subjects to the low shared-attitudes-and-values roommate condition. Those students are paired for room assignment where there is only 20% agreement in their attitudes/values responses. The director randomly assigns another 40 students to the medium shared-attitudes-and-values roommate condition. Each of these 20 roommate pairs had approximately 50% agreement in their attitudes/values responses. The remaining 40 students were assigned to the high shared-attitudes-and-values condition. These 20 pairings were made so each had approximately 80% agreement on the attitudes and values survey.

All 120 students completed a roommate likability questionnaire at the end of the first semester. The questionnaire was designed by the researcher to assess several aspects of how roommates were liked. An average roommate likability score was computed for each of the 60 roommate pairs. This average roommate likability was the score for each roommate pair and constituted the data analyzed by the residence hall director.

Analyses of the data by the Director of Residence Life did not support the old adage that opposites attract. That is, the high shared-attitudes-and-values group had a higher average roommate likability score than did the medium shared-attitudes-and-values group. This medium group had a higher average roommate likability score than did the low shared-attitudes-and-values group. The differences among these three means were larger than expected, by chance, at the .001 level of significance ($p < 0.001$). Fifty-four percent of the variability in roommate likability scores was accountable by knowing the proportion of shared attitudes and values of the roommate pairs (eta-squared = .54). A 99% confidence interval indicated that the mean population roommate likability would be 18 to 27 points higher if roommates were paired with high rather than low shared-attitudes-and values.

VIGNETTE 3

A developmental psychologist interested in understanding variability in children's self-esteem studied the relationship between parenting style and a child's social-self-esteem. She included in her study only children whose parents had not been divorced and were still living together with their children. Using the fifth-grade children in the school district nearest her university, the psychologist administered a social-self-esteem scale to the children and a parenting-behavior inventory to their parents. The parenting-behavior inventory permits the classification of parents into different parenting-style groups. The psychologist identified 20 sets of parents whose parenting style was democratic, 20 sets of parents whose parenting style was authoritarian, and 20 sets of parents whose parenting style was classified as permissive. Another 45 sets of parents were not used in the study because their parenting-behavior inventory responses did not permit a clear classification into one of these three parent types. The social-self-esteem score for each of the 60 fifth graders in the study was determined with an instrument developed by the research based on research by Coopersmith and Harter. The researcher constructed the social-self-esteem scale for use in an earlier study in which she found her scale correlated highly with Harter's social-self-esteem.

Analyses of the data by the developmental psychologist indicated that the democratically-reared children's average social-self-esteem was higher than the average for the permissively-reared children and the average for the permissively-reared children was higher than that of the authoritarian-reared children.

The differences among these three means were larger than expected by chance at the .05 level of significance ($p < 0.5$). Fifty-nine percent of the variability in social-self-esteem scores was accountable by knowing which parenting style was used (eta-squared = .59). A 95% confidence interval indicated that a democratically-reared population would be somewhere between 20 to 30 points higher in a social-self-esteem than an authoritarian-reared population of fifth graders.

VIGNETTE 4

A personnel psychologist working for a fast-food chain was interested in accounting for variability in the success of employees. She studied the relationship between the achievement motivation level of employees and their success on the job at the company's 8,000 franchises. She arranged for the company to give an achievement motivation scale to all applicants. The personnel psychologist took a random sample of 50 applicants who scored between the 15th and the 25th percentiles on the achievement motivation scale, a random sample of 50 applicants who scored between 45th and the 55th percentiles, and a random sample of 50 applicants who scored between the 75th and the 85th percentiles on the achievement motivation scale. All 150 of these subjects were hired. The franchise managers were not aware of the applicants' scores on the achievement motivation scale.

After six months on the job, an employee success score was obtained for each of the 150 subjects. Success points were earned for punctuality, dependability, following orders, and willingness to work hard. The success score was the sum of points earned by each employee.

Analyses of the data by the personnel psychologist indicated the 15-25% achievement motivation group had lower average employee success score than the 45-55% achievement motivation group and the 45-55% achievement motivation group had a lower average employee success score than did the 75-85% achievement motivation group. The differences among these three means were larger than expected at the .05 level of significance ($p < .05$).

Eleven percent of the variability in success scores was accountable by knowing the employees' achievement motivation score (eta squared = .11). A 95% confidence interval indicated that the population of high achievement motivation applicants would have an average success score between 2 and 18 points higher than the population of low achievement motivation applicants.

VIGNETTE 5

A clinical psychologist studied the relationship between therapeutic approach and success in overcoming acrophobia (a fear of high places). Using the 1990 census data, he sent a questionnaire to a large random sample of individuals who were between 18 and 50 years old. One question asked respondents to rate their level of fear of being in high places. The psychologist randomly sampled 45 individuals who gave the highest rating on this question and invited them to participate in a treatment program for their acrophobia. He randomly assigned 15 subjects to a psychoanalytic treatment program plan, 15 subjects to a behavior modification treatment plan, and 15 to a cognitive therapy treatment plan.

After six months of treatment, each subject was tested for the strength of his/her acrophobia response. This testing was carried out by clinical psychologists who were unaware of the treatment program to which the subject had been assigned. The subject's acrophobia score was the sum of acrophobia points awarded by two clinical psychologists.

Analyses of the data by the clinical psychologist indicated the Cognitive Therapy Group had a lower average acrophobia score than did the Behavior Modification Group. The Behavior Modification Group had a lower average acrophobia score than did the Psychoanalytic Group. The differences among these three means were larger than expected by chance at the .05 level of significance ($p < .05$).

Thirty-five percent of the variability in acrophobia scores was accountable by knowing the therapy treatment given to the subjects (eta-squared = .35). A 95% confidence interval indicated that the population mean under the cognitive therapy treatment is somewhere between 8 and 15 points lower than the population mean under behavior modification treatment.

VIGNETTE 6

A social psychologist studied the relationship between the physical distance of two people and a person's rated level of social comfort in a conversational context. She included in her study only females between the ages of 18 and 35 who were registered voters in Pennsylvania. The male, with whom each subject had a conversation, was 28 years old and moderately physically attractive. The psychologist randomly sampled 42 females from a population of 18 to 35-year-old PA-registered female voters. Each subject was transported to the site of the study and escorted to a room with furniture in a lounge arrangement. She was met by the researcher's 28-year-old male accomplice who explained that he was conducting interviews about voters' views on several political issues. The lounge was arranged the same way for each interview. After the subject was seated on a sofa, the experimenter sat down and began the 15-minute-long interview. Fourteen subjects were randomly assigned to each of three interviewing conditions. The experimenter sat within the personal space (1 to 1-1/2 feet) of subjects in Group A. He sat an informal distance (approximately 3-1/2 feet) from subjects in Group B and at a formal social distance (approximately 5-1/2 feet) from subjects in Group C.

At the conclusion of the interview, the social psychologist introduced herself to the subject, explained that the interviewer was a trainee, and asked the subject to rate several aspects of the interviewer's performance. The questions asked the subject to rate different aspects of her level of comfort during the interview using a seven-point scale for each question. The sum of these ratings was the score for each subject. Analyses of the data by the social psychologist indicated subjects seated 1-1/2 feet from the interviewer (Group A) reported an average comfort that was greater than those seated 3-1/2 feet from the interviewer (Group B) and this average comfort for Group B was higher than the average comfort reported by the subjects seated 5-1/2 feet from the interviewer (Group C). The differences among these three means were larger than expected by chance at the .005 level of significance ($p < .005$). Twenty-nine percent of the variability in comfort ratings was accountable by knowing the distance of the interviewer from the subject (eta-squared = .29). A 95% confidence interval indicated that the population of Pennsylvania females sampled would rate the interviewer 0.4 to 2.2 points higher at the one and one-half foot distance than at the five and one-half foot distance.

BEST COPY AVAILABLE

VIGNETTE 7

A social psychologist interested in environmental variables that affect peoples' moods studied the relationship between the type of movie a person has just seen and self-reported feelings of depression. She gathered her data at a large mall that had 10 movie theatres at a single location in the mall. As individuals exited one of the theatres, they were greeted by the researcher and offered a certificate for a free soda or lemonade if they would be willing to complete a mood questionnaire. If willing, the subject was asked to return to meet the researcher after a 15 minute interval during which they could enjoy their free drink. One-third of the 72 subjects had just seen a movie classified as a comedy, an other third had exited a drama, and the remaining 24 subjects were at movies advertised as thrillers.

When returning from consuming their free drink, each subject was asked to complete a self-report depression scale devised by the researchers to assess a subject's immediate mood state. The score for each subject was the sum of the items on the scale with a high score reflecting higher reported feelings of depression.

Analyses of the data indicated that the comedy movie group had a higher average depression score than the drama movie group and the drama group had a higher depression score than the thriller movie group. The difference among these means were larger than expected by chance at the .05 level of significance ($p < .05$). Twenty-two percent of the variability in depression scores was accountable by knowing the type of movie subjects had just witnessed (eta-squared = .22). A 95% confidence interval indicated that the population mean for comedy audiences would be 1.7 to 4.6 points higher than the population mean for the thriller audiences.

VIGNETTE 8

A researcher was interested in factors that relate to people's belief in psychic or other occult phenomena. In one of his studies, he asked students in several General Psychology sections to examine a list of 100 movies and to check which ones they had seen during the last two months. Included in the list were several films on the occult such as "The Omen", "Scanners", and "Carrie". Using this information, the researcher selected 25 students who reported seeing no occult videos, 25 students who reported seeing two or three of the occult films, and 25 students who reported seeing five or six of the occult films on the list. These 75 students were invited to a show put on by a magician who was billed as a traveling psychic. The magician demonstrated her "psychic powers" by reading books while blindfolded, finding hidden objects, and causing several objects to burst into flame. After the performance, all 75 students were given a survey that asked them to rate how confident they were that each of the events they witnessed was the result of psychic powers. The score for each subject was the sum of the confidence ratings across events witnessed.

Analyses of the data indicated that the group that saw two or three occult films had a higher average psychic-power rating than those who saw five or six occult films. The five-or-six-films group gave a higher mean psychic power rating than those who saw no occult films. The differences among these three means were larger than expected by chance at the .01 level of significance ($p < .01$). Forty-two percent of the variability in confidence-in-psychic-power ratings was accountable by knowing the movie group to which subjects belonged (eta-squared = .42). A 99% confidence interval indicated that the population viewing two or three occult videos would be 13-21 points higher in their confidence mean than the population viewing no occult films.

VIGNETTE 9

A clinical psychologist studied the relationship between therapy follow-up procedures and the success that obese patients have in losing weight. She used 120 patients who were being treated for obesity by the Health Maintenance Organization (HMO) where she was the staff psychologist. All 120 patients first participated in a 10-week behavior modification plan to reinforce proper eating and exercise regimens. She randomly assigned 40 of these individuals to follow-up Group A. They came to the clinic for a once-a-week weigh-in for 10 more weeks. The 40 patients randomly assigned to follow-up Group B also came to the clinic for a weekly weigh-in for the second 10 weeks. However, in addition to the weigh-in, Group B subjects met with the therapist each week to review progress on their behavior modification plan. The 40 Group C subjects were paired up with support partners who check with each other daily on their success in staying on the prescribed diet and exercise regimens. They also came to the clinic each week for weigh-in.

At the end of 15 weeks, after the behavior modification training sessions, a final weigh-in was made and compared to the initial weigh-in. The 25-week weight change was the score recorded for each subject.

Analyses of the weight-change data by the clinical psychologist indicated the support partner Group C, had a greater average weight loss than the average for the therapist follow-up Group B. The average weight loss for his therapist follow-up group was greater than for the weigh-in-only Group A.

The differences among these three means were larger than expected by chance at the .05 level of significance ($p < .05$). Eleven percent of the variability in weight change scores was accountable by knowing the follow-up condition to which the subjects had been assigned (eta-squared = .11). A 95% confidence interval indicated that the population mean weight loss is somewhere between 15 and 25 pounds greater in the support-group follow-up condition than in the weigh-in-only follow-up condition.

VIGNETTE 10

A sport psychologist was interested in factors that affect ratings given by diving judges. He studied the relationship between information about a diver's team standing and the rating of a dive, given by a judge. He used videos of male divers and subjects who had judged NCAA Division II diving meets in the last year. The psychologist randomly sampled 30 judges from all individuals who served as an NCAA Division II diving judge. All judges observed the same videotape, showing dives by 10 NCAA Division II men. As each diver approached the board, his name was announced along with his standing as his team's first, second, third, fourth, or fifth-place diver. The judges were randomly assigned to one of three groups. Judges in Group A were told that the fourth diver was his team's first-ranked diver. Group B judges were told that the fourth diver was his team's third-ranked diver. Group C judges were told that the fourth diver was his team's fifth-ranked diver. These three groups of judges were compared on the ratings given to the fourth diver on the video. The judges' ratings of the fourth diver was on the standard 10-point rating scale used in diving meets.

Analyses of the data by the sport psychologist indicated that the average rating given to the fourth diver by Group A was higher than the average given by Group B and the average rating given by Group B was higher than the average given by Group C. The differences among these three means were larger than expected by chance at the .05 level of significance ($p < .05$). Eighteen percent of the variability in ratings of diver number four was accountable by knowing what the judges were told about the diver's rank on his team (eta-squared = .18). A 95% confidence interval indicated that the population of judges would rate a first-within-team diver 0.1 to 0.7 points higher than they would a fifth-within-team diver.

VIGNETTE 11

A developmental psychology faculty member was interested in helping her students better understand and retain course concepts. She interviewed several of the very best students and those who appeared to be struggling the most in her developmental courses. It appeared from these interviews that many of the students who were quite successful were using a self-referencing approach. That is, they would relate each concept to their own life experiences and, thus, have an example for each concept. Many who seemed to be struggling the most seemed to have no strategy for learning. Many from both groups used highlighting as they read the book. In order to systematically examine the effectiveness of self-referencing and highlighting, the faculty member randomly sampled 210 students at her university to be in a learning study. One-third of the subjects were given instructions and practice on how to use self-referencing in the study process. The second randomly assigned group was told that the purpose of the study was to identify effective study strategies. These 70 Group B subjects were told to do their best in studying the assigned material and that they would be interviewed later for the researcher to learn what study method was used. The third randomly assigned group of subjects (Group C) were told to highlight the important sections of the chapter and to review the highlighted sections after reading and highlighting. All subjects were given the same chapter to study and the same amount of time to master it. Three weeks after studying the chapter, all subjects took the same test. Their success was measured by the percent correct on the test.

The analyses of the data indicated that the self-referencing Group A had a higher mean percent correct than did the highlighting Group C. Group C had a higher mean percent correct than did the do-your-best Group B. The differences among these three means were larger than expected by chance at the .01 level of significance ($p < .01$). Fifty-eight percent of the variability in the percent correct grades was accountable by knowing the study strategy group to which the subject belonged (eta square = .58). A 99% confidence interval indicated that the mean percent correct for the population of students at the university would be 9.1 to 16.3 points higher using self-referencing compared to the do-your-best study strategy.

VIGNETTE 12

An industrial psychologist studied the relationship between participation in a physical fitness program and health-related absences from work. She was able to study this relationship in a large corporation where executives were concerned about the excessive number of health-related absences by their employees. Using one of the corporation's plants, the industrial psychologist identified 60 employees who had been given brochures about the importance of physical fitness along with an announcement of the opening of the company's new fitness center. The brochure included several suggestions for becoming more physically fit. The benefits emphasized in the brochures were happiness, less stress, and greater sexual activity. No mention was made of reduced work absences. Twenty of the subjects did not make use of the company's new fitness center. Another 20 subjects signed up for and participated in the 15 minutes of exercise per working day program at the company's new fitness center. The remaining 20 subjects signed up for and participated in the 30 minutes of exercise per working day program at the company's new fitness center. The exercise time was part of the employee's work day for the latter two groups for a period of 12 months.

One month after the fitness program began, health-related absences for each of the 60 employees were recorded for a period of one year. This count of the number of health-related absences was the score given to each subject. Analyses of the data by the industrial psychologist indicated the brochure-only group had more health-related absences than the 15-minutes-of-exercise-per-day group and the 15-minutes-per-day group had more health-related absences than the 30-minutes-of-exercise-per-day group. The differences among these three means were larger than expected by chance at the .01 level of significance ($p<.01$). Fifty-three percent of the variability in health-related absences was accountable by knowing the subject's degree of participation in using the company's fitness center (eta-squared=.53). A 99% confidence interval indicated that the population of 30-minute-per-day exercisers would average between nine to 16 fewer absences in a year than the no-exercise population.

VIGNETTE 13

A psychometrician studied the relationship between the advice given to students about test-taking strategy and students' success on a multiple-choice test. He used sections of his General Psychology course to explore this relationship. He randomly assigned each of the 180 students in his class to one of three identical testing rooms for the final exam. Before the 60 students in each room began the final exam, the professor offered some advice. Group A was told the advice given by most faculty, counselors, and test corporations: "Answer all questions, stick with your first choice on multiple-choice questions, and change answers only if you are absolutely certain your initial choice is incorrect." Group B subjects were told to answer each of the multiple-choice questions and then be liberal about changing answers as they review the test because such changes are likely to lead to more correct answers rather than wrong ones. Group C is simply told there will be no penalty for guessing, and they should be sure to choose an answer for every multiple-choice question.

The score for each student was the number correct on the multiple-choice final exam.

Analyses of the data by the psychometrician indicated that Group B (change answers liberally) earned a higher average on the final exam than Group A (stick with the initial choice), and Group A had a higher average on the final exam than Group C (no penalty for guessing). The differences among these three means were larger than expected by chance at the .05 level of significance ($p<.05$). Six percent of the variability in final exam scores was accountable by knowing which instructions were given to the students (eta-squared=.06). A 99% confidence interval indicated that the population mean of students would be between .09 to 5.3 points higher if advised to change multiple choice answers liberally rather than staying with their initial choice.

VIGNETTE 14

An educational psychologist was interested in factors affecting test performance. An earth science faculty colleague studying air pressure changes on plants had constructed a room fitted with large air compressors that allowed researchers to vary the level of air pressure in the room. The educational psychologist had her 87 introduction to the educational psychology students take their mid-term exam in the special air pressure room. She randomly assigned one-third of her students to each of three testing times. Those in Group A had an air pressure of 28.5 in the room during the test. Group B students had a pressure reading of 29.75 during their test and Group C had the air pressure set at 31 during their test. As the air pressure changed to these settings, a sound system installed in the room was used to simulate the sounds of rain, wind, and occasional thunder. The same sound track was used for all three groups. The percent correct on the same exam given to all three groups was the score assigned to each student.

The analyses of the data indicated that the high-pressure Group C had a higher mean percent correct on the mid-term than the medium-pressure Group B and that Group B had a higher mean grade than the low-pressure Group A. The differences among these three means were larger than expected by chance at the .05 level of significance ($p<.05$). Eight percent of the variability in the percent correct on the mid-term grades was accountable by knowing the air pressure in the room for each student taking the test (eta-squared=.08). A 95% confidence interval indicated that percent correct on the mid-term exam would be 0.2 to 3.1 for the population of students if tested under the high rather than low air pressure conditions.

VIGNETTE 15

A developmental psychologist was interested in variables that account for differences in elementary school children's self-efficacy. In one of her studies, she examined the relationship between the orientation of youth soccer coaches and a fifth grader's sport self-efficacy. She first administered a questionnaire to all volunteer coaches in the American Youth Soccer Organization (AYSO). It assessed the coach's attitudes about issues such as everyone plays, and the primary program goals as winning, development of fitness, fun, social skills, or soccer skills. Based on their responses, she randomly selected 35 coaches who were oriented primarily to keeping youth physically fit (Group A), 35 coaches who were oriented primarily to skill development and winning (Group B), and 35 coaches who were oriented primarily to fun and social skill development (Group C).

Three-quarters of the way through the Fall season, the psychologist assessed the self-efficacies of each fifth and sixth grade child coached by these 105 coaches. The following sport self-efficacy item was analyzed in this study:

> "When I compare my abilities with those of my classmates in the areas of mathematics, reading, art, sports, science, music, and social studies, I consider sports to be: _(1) My weakest ability area, _(2) one of my weakest ability areas, _(3) a below average ability area for me, _(4) an average ability area for me, _(5) an above average ability area for me, _(6) one of my strongest ability areas, or _(7) my strongest ability area." The score for each coach was the average sport self-efficacy score for his/her team.

Analyses of the data indicated that Group C coaches had players with a higher sport self-efficacy than Group B whose players had a higher sport self-efficacy than Group A. The differences among these three means were larger than expected by chance at the .05 level of significance ($p<.05$). Seven percent of the variability in sport self-efficacy was accountable by knowing the coaches' philosophy (eta-squared=.07). A 95% confidence interval indicated that the population of children coached by fun-oriented coaches would have a sport self-efficacy that is 0.1 to 2.9 points higher than the population coached by fitness-oriented coaches.

VIGNETTE 16

A sport psychologist specializing in motivational factors affecting athletic performance studied the relationship between hours of practice per week and athletic performance. He contacted track coaches in Division II colleges and universities in Pennsylvania for help in recruiting subjects. Fifteen coaches agreed to keep a record of the hours of practice per week for four of their male 400-meter runners. These records were sent to the sport psychologist three quarters of the way through track season. Because of his concern about perfect precision in record keeping, he simply divided the 60 track-team members into three groups. Group A practiced between 8 and 12 hours (Average=10) per week, Group B practiced between 13 and 17 hours (Average=15) per week, and Group C practiced 18 or more hours (Average=20) per week.

All 60 runners were invited to a special 400-meter track event at the sport psychologist's university. The running time in that event was the score for each subject in the study.

Analysis of the data by the sport psychologist indicated that the group that practiced an average of 15 hours per week (Group B) had a faster average running time than those averaging 10 hours of practice per week (Group A), and Group A had a faster running time average than the group practicing an average of 20 hours per week (Group C). The differences among these three means were larger than expected by chance at the .01 level of significance ($p<.01$). Thirty-nine percent of the variability in running times was accountable by knowing the hours-of-practice group to which the subject belonged (eta-squared=.39). A 99% confidence interval indicated that running time in the 400 meter event would be .09 to 1.3 seconds faster for the population practicing about 15 hours per week than for the population practicing approximately 20 hours per week.